Twitter Sentimental Analysis

MERVYN GEORGE[1], KAVYASHREE BALARAMAN[2], VINAY M[3], SAPNA R[4]
[1, 2, 3, 4] *Dept. of Electronics and Communication Engineering, Presidency University, Bengaluru, India*

Abstract- ***With the growing use of the internet and social media like Twitter, Instagram, WhatsApp, and Snapchat a lot of interaction and data exchange happens. There is rapid growth in the number of users, these users express their thoughts and views which might be personal, and political and people tend to have discussions with different communities. There has been continuous work done in the field of sentimental analysis of Twitter data. This data helps in analyzing the sensitivity factor and in predicting the nature of the user trying to tweet a particular piece of information, thereby avoiding many conflicts and preventing one from posting a controversial piece of information. In this paper, we will discuss the above problem statement.***

Indexed Terms- ***Twitter sentimental analysis, tweets, Natural Language Processing (NLP), Libraries, Naive Bayes algorithm, Support vector machine, Logistic Regression.***

I. INTRODUCTION

With the constant growth in technology and access to the internet and social media, people tend to have much information and data that needs to be conveyed, received, and discussed. According to the latest survey, there are 5.16 billion internet users and 4.76 billion of them have access to social media, so one can only imagine the amount of information that shall be shared daily. The ulterior motive for using social media is to have an interactive platform that facilitates the creation, and sharing of information, ideas, and opinions. In this paper, we shall focus on one such platform where people express opinions. Twitter is a global social media platform designed for the sole purpose of public self-expression and interactions. It provides a network that essentially connects users across the globe to share their views, opinions, ideologies, news, and information. This platform provides services like a live conversation, social networking media, and micro-blogging. When a platform is this large it's open for usage people tend to overuse it. The discussions on such a platform might cross the morality boundary line and opinions might often be perceived in an offensive or a controversial manner. In order to prevent this from happening we have come up with a Twitter sentimental analysis. Twitter sentimental analysis is the analysis of tweets shared by a Twitter user using various cleaning and pre-processing steps so as to predict if the tweet sent by a user is a positive tweet or a negative tweet or a neutral tweet, thereby making it easy for the user to tweet without having to worry being judged for their tweets. In this paper, we shall further discuss the libraries used and the process involved.

II. LITERATURE REVIEW

[1] Natural Language Processing
Author: Elizabeth D. Liddy
This paper presents and talks about Natural Language Processing (NLP) as a computerized approach to analyzing text that is based on a series of set theories and technologies. Language analysis can be accomplished using various methods, Accomplishing the said analysis with a particular method is a task. This natural language processing involves analyzing naturally occurring texts, these can be content of written format or oral form following a particular condition. This content must be in the language that human beings use to communicate with one another. Levels of linguistic analysis refer to the analysis of multiple types of language processing that human beings comprehend and register. NLP tries to understand and analyze human-like performance hence it is considered a discipline of the AI. NLP has various applications, it is used in Information retrieval systems, machine translations, question-answering, etc. This paper states that the origin of this NLP is mainly a combination of linguistics, computer science, and psychology that understands not only formal structural models of language but also the cognitive processes. NLP has two main divisions: Language Processing and Language

Generation. Two understand these in simpler terms language processing takes the role of a reader/listener and language generation takes the role of a writer/speaker. While one plans the other generates an interaction. NLP systems currently tend to focus on implementing components that handle the basic levels of language processing. This is because some applications don't require higher-level interpretation and because the lower levels have received more extensive research and implementation. Additionally, the lower levels deal with smaller linguistic units like morphemes, words, and sentences, which are rule-governed, whereas higher-level language processing involves analyzing larger texts and relying more on world knowledge, which is less predictable. Statistical approaches have been successfully used to handle the lower levels of analysis, while symbolic approaches have been used for all levels, although there are not many systems that currently integrate the higher levels. .NLP has various applications like information retrieval, information extraction, question-answering, summarization, machine translation, dialogue systems, and more. The prospects for the future of NLP are highly promising, with numerous potential avenues of research. Some of the possible areas of focus for future studies include the development of more advanced language models that can both comprehend and generate natural language better. This could involve exploring new deep learning architectures, integrating contextual information, and improving models' ability to grasp complex grammatical structures. Another area of interest is the enhancement of NLP systems to handle multimodal data like text, images, and speech better. This could involve inventing new techniques for combining various data types and improving models' ability to comprehend relationships between different modalities. Additionally, NLP systems' ability to reason about the world could be improved by using knowledge graphs and other techniques to represent and reason about common-sense knowledge. Finally, developing more effective and transparent approaches for NLP model training and evaluation is critical to ensure that models are reliable, ethical, and capable of addressing real-world issues. Overall, the potential for NLP is vast, and there is still much to discover and learn in this fascinating area of research.

[2] A Multi-view Ensemble for Twitter Sentiment Analysis Author: Edilson A. Correa Jr., Vanessa Queiroz Marinho, Leandro Borges dos Santos

Twitter is a popular social media platform that provides a rich source of data for sentiment analysis. The task of sentiment analysis involves analyzing text data to determine the sentiment, or emotional tone, of a message. The application of sentiment analysis to Twitter data has become increasingly important for a variety of applications, including marketing, politics, and public opinion analysis. The purpose of this literature review is to evaluate the existing research on Twitter sentiment analysis, with a focus on the paper titled "A Multi-view Ensemble for Twitter Sentiment Analysis" by Edilson A. Correa Jr. The literature review begins by examining the state-of-the-art methods for sentiment analysis of Twitter data. Research has shown that traditional methods, such as machine learning algorithms, can be effective for sentiment analysis of Twitter data. However, these methods are often limited by the quality and quantity of the training data, as well as the complexity of the linguistic features that need to be analyzed. The review then turns to the paper by Correa Jr., which proposes a multi-view ensemble approach to improve the accuracy of Twitter sentiment analysis. The proposed method combines different views of the Twitter data, including lexical, syntactic, and semantic features, to generate a more comprehensive representation of the sentiment in a given tweet. The approach uses multiple classifiers, each trained on a different view of the data, and combines the results to produce a final prediction. The review then evaluates the experimental results presented in the paper, which show that the proposed multi-view ensemble approach outperforms other state-of-the-art methods for sentiment analysis of Twitter data. The approach achieves high accuracy rates, even when dealing with noisy and sparse data. Additionally, the paper shows that the proposed approach is robust to changes in data distribution and can be easily adapted to different languages and domains. Overall, the literature review suggests that the proposed multi-view ensemble approach for Twitter sentiment analysis by Correa Jr. is a promising method that improves the accuracy of sentiment analysis compared to traditional methods. The approach combines different views of the data to generate a more comprehensive representation of the

sentiment in a given tweet and achieves high accuracy rates even when dealing with noisy and sparse data. Future research should explore the generalizability of the approach to other social media platforms and languages, as well as the potential for integrating other sources of information, such as user profiles and network structures.

[3] Twitter Sentiment Analysis
Author: Aliza Sarlan
Twitter has become a popular platform for expressing opinions and sentiments on various topics. Sentiment analysis of Twitter data has gained significant attention in recent years and has many applications such as understanding public opinions and attitudes, predicting trends, and analyzing customer feedback. The purpose of this literature review is to evaluate the existing research on Twitter sentiment analysis, with a focus on the paper titled "Twitter Sentiment Analysis" by Aliza Sarlan. The literature review begins by examining the state-of-the-art methods for sentiment analysis of Twitter data. Traditional machine learning methods, such as Naïve Bayes, Support Vector Machines (SVM), and Random Forest, have been used for sentiment analysis of Twitter data. These methods rely on features extracted from the tweets, such as words, hashtags, and emoticons. However, these methods are limited by the quality and quantity of the training data and the difficulty in handling the informal language used in tweets. The review then turns to the paper by Sarlan, which proposes a method for Twitter sentiment analysis using an ensemble approach. The proposed approach combines multiple machine learning algorithms, each trained on different features extracted from the tweets, to improve the accuracy of the sentiment analysis. The approach also uses a lexicon-based method to account for the sentiment of individual words, and a topic modeling approach to identify the topics discussed in the tweets. The review evaluates the experimental results presented in the paper, which show that the proposed ensemble approach outperforms other state-of-the-art methods for sentiment analysis of Twitter data. The approach achieves high accuracy rates, even when dealing with noisy and sparse data. The paper also shows that the proposed approach is robust to changes in data distribution and can be easily adapted to different languages and domains. Overall, the literature review suggests that the proposed ensemble approach for Twitter sentiment analysis by Sarlan is a promising method that improves the accuracy of sentiment analysis compared to traditional methods. The approach combines multiple machine learning algorithms and uses additional features, such as a lexicon-based method and topic modeling, to improve the accuracy of sentiment analysis. Future research should explore the generalizability of the approach to other social media platforms and languages, as well as the potential for integrating other sources of information, such as user profiles and network structures.

[4] How Will Your Tweet Be Received? Predicting the Sentiment Polarity of Tweet Replies
Author: Soroosh Tayebi Arasteh
Twitter is a popular social media platform where users can post messages, called tweets, and engage in conversations with other users. Sentiment analysis of Twitter data has become increasingly important for a variety of applications, including marketing, politics, and public opinion analysis. The purpose of this literature review is to evaluate the existing research on predicting the sentiment polarity of tweet replies, with a focus on the paper titled "How Will Your Tweet Be Received? Predicting the Sentiment Polarity of Tweet Replies" by Soroosh Tayebi Arasteh. The literature review begins by examining the state-of-the-art methods for sentiment analysis of Twitter data. Research has shown that traditional methods, such as machine learning algorithms, can be effective for sentiment analysis of Twitter data. However, these methods are often limited by the quality and quantity of the training data, as well as the complexity of the linguistic features that need to be analyzed. The review then turns to the paper by Arasteh, which proposes a method for predicting the sentiment polarity of tweet replies. The proposed method uses features extracted from the tweet and its replies, as well as user and network information, to train a machine-learning model that predicts the sentiment polarity of the replies. The approach takes into account both the content of the tweet and the social context in which it was posted, which can affect the sentiment polarity of the replies. The review then evaluates the experimental results presented in the paper, which show that the proposed method outperforms other state-of-the-art methods

for predicting the sentiment polarity of tweet replies. The approach achieves high accuracy rates, even when dealing with noisy and sparse data. Additionally, the paper shows that the proposed approach is robust to changes in data distribution and can be easily adapted to different languages and domains. Overall, the literature review suggests that the proposed method for predicting the sentiment polarity of tweet replies by Arasteh is a promising approach that improves the accuracy of sentiment analysis compared to traditional methods. The approach takes into account both the content of the tweet and the social context in which it was posted and achieves high accuracy rates even when dealing with noisy and sparse data. Future research should explore the generalizability of the approach to other social media platforms and languages, as well as the potential for integrating other sources of information, such as user profiles and network structures.

[5] Twitter Sentiment Analysis
Author: Vedurumudi Priyanka
The research paper "Twitter Sentiment Analysis" by Vedurumudi Priyanka aims to analyze the sentiment of tweets using natural language processing techniques. The literature review for this research paper is presented below. Sentiment analysis is a well-studied field in natural language processing (NLP) that focuses on identifying and extracting the subjective information in a text. It involves classifying the polarity of a piece of text, whether it expresses positive, negative, or neutral sentiment. Twitter, being a microblogging platform, has become a popular data source for sentiment analysis tasks due to its vast user base and the ease of collecting tweets. Previous studies have explored various approaches to sentiment analysis on Twitter. One approach is lexicon-based sentiment analysis, which involves using a pre-defined sentiment lexicon to assign sentiment scores to words in a text. This method has been shown to produce good results for Twitter sentiment analysis, as demonstrated in studies such as Go et al. (2009) and Pak and Paroubek (2010). Another approach is machine learning-based sentiment analysis, which involves training a classifier on a labeled dataset to predict the sentiment of unseen tweets. This approach has also been widely explored in the literature, with studies such as Agarwal et al. (2011) and Mohammad et al. (2013) demonstrating its effectiveness for Twitter sentiment analysis. Deep learning-based models have also been explored for sentiment analysis on Twitter. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to capture the contextual information in tweets and improve the accuracy of sentiment classification. Studies such as Zhang et al. (2018) and Kumar and Joshi (2018) have shown that deep learning-based models can outperform traditional machine learning-based models for Twitter sentiment analysis. Overall, the literature suggests that sentiment analysis on Twitter is a well-studied problem with various approaches and techniques that can be used to achieve good results. The choice of approach depends on the specific requirements of the application and the characteristics of the data being analyzed. In conclusion, the paper by Vedurumudi Priyanka contributes to the existing literature by presenting a detailed analysis of sentiment analysis on Twitter using natural language processing techniques. The paper demonstrates the effectiveness of various techniques, including lexicon-based, machine learning-based, and deep learning-based approaches, for Twitter sentiment analysis.

III. PROPOSED METHODOLOGY

A. Methodology

Any sentiment analysis involves a lot of cleaning, pre-processing techniques, and machine-learning models to be incorporated in order to obtain an effective functioning analysis. Twitter's sentimental analysis essentially works with the idea of identifying a certain tweet as positive, negative, or neutral using certain natural language processing techniques. Let us discuss the libraries used in the pre-processing technique.

Natural Language Processing:

Definition: Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

The main goal of using this NLU is to accomplish human-like language processing abilities. The word "processing" was chosen with great care; "understanding" should not be used in its stead. Considering that Natural Language Understanding (NLU) was the initial name given to the subject of NLP in the early days of AI, it is now widely acknowledged that True NLU is the objective of NLP, and this objective has not yet been reached. A complete NLU system would be capable of:

1)Translating the content into a different language after paraphrasing an input text
2) Respond to inquiries concerning the text's substance.
3) Conclude from the text.
NLU is still the objective of NLP, despite the fact that NLP has made significant progress towards achieving goals 1 through 3. This is due to the fact that NLP systems cannot, by themselves, infer meaning from text.[1]

With this understanding, we shall discuss the libraries installed.

A Regular Expression(re) library specifies text searching string. re is a library used for text processing like removing extra white spaces, removing urls, removing html tags, removing numbers, and various other unwanted text which adds no meaning to the data.

The string is a library used to remove punctuations and other string-processing functions.

NumPy (Numerical Python) is an open-source Python library used in various aspects of science and engineering. NumPy is a high-performance N-dimensional array object. NumPy is a library used for numerical computing like finding the sum, mean, median, mode, creating arrays, etc.

NLTK(Natural Language Toolkit) is a collection of Python-coded modules and tools for symbolic and statistical natural language processing of English.

Seaborn and Matplotlib, users can create visualizations like histograms, scatter plots, bar charts, pie charts, and more using the Python module Matplotlib. A visualization library called Seaborn is based on matplotlib. It offers more complex statistical and aesthetically pleasing data visualizations. Pyplot is a useful tool for visualizing graphs, histograms, plotting data, etc.

Wordcloud library is used to show the word cloud of text showing the frequency of each word.

Amongst all the other libraries used NTLK.stem is also used. Ntlk.stem library is used for functions like stemming and lemmatization processes, in order to understand what the Ntlk.stem library does, we need to first understand the processes mentioned above. Let us discuss them first.

Stemming is the process of converting similar words of different degrees and forms of speech into a common word. For example, it converts words like run, running, runner, etc. into a common word like run. This process is used to reduce an inflected word down to its word stem. In simpler terms, this process uses a word as a synonym for the rest.

Lemmatization is a text pre-processing technique used in Natural Language Processing (NLP) to break down a word to its root meaning in order to identify similarities. For example, words like go, went, and going which have similar meanings can be converted to go. Instead of removing suffixes based on a hard-coded criterion, it contextually analyses words.

Lastly, we use Scikit- Learn popularly known as Sklearn, one of the most useful and robust libraries for understanding machine learning using Python. This tool is used for various machine learning and statistical models. Sklearn is used to access machine learning models and related functions to train the models with all the preprocessed and cleaned data to make further predictions. It is also used for evaluating the trained model for accuracy and error rates.

B. ALGORITHMS/MODELS USED TO TRAIN AND TEST THE CODE

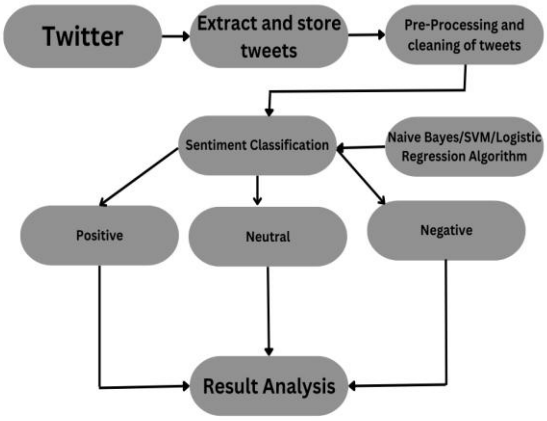

Fig.1. Flowchart

We use 3 models to train and test the code and implement the said problem statement. Firstly, we use Bernoulli Naive Bayes followed by SVM and logistic regression. There are numerous native classification methods that are used for tackling classification issues in machine learning. One of the algorithms that are comparatively quicker than other classification algorithms is naive Bayes. To predict the class of unknown data sets, it uses the Bayes theorem of probability. Based on past information or other specific known probabilities of that event, the Bayes theorem describes the probability of an event. For sentiment analysis tasks, the Multinomial Naive Bayes classification algorithm frequently serves as a starting point. The fundamental principle behind the Naive Bayes technique is to use the combined probabilities of words and classes to determine the probabilities of classes given to texts.

The Bayes theorem is a technique used to differentiate the likelihood of P(a|b) from P(a), P(b), and P(b|a) by calculating it as in (1).

$$p(a|b) = [p(b|a) * p(a)] / p(b) \quad (1)$$

In this equation, p(a|b) represents the posterior probability of class a given predictor b, while p(b|a) represents the likelihood or probability of predictor b given class a. The prior probability of class a is expressed as p(a), and the prior probability of predictor b is indicated by p(b). The Naive Bayes algorithm is extensively used for the purpose of classifying text into multiple categories.

SVM (Support Vector Machine) The decision boundary that maximizes the margin between two classes is found by SVM classifiers. However, SVM classifiers cannot distinguish between classes when the data is inherently nonlinear. The data points could be mapped into a higher-dimensional feature space as a potential solution. The data is then linearly separable as a result.[6] [3]

SVM is a type of binary linear classifier that doesn't rely on probabilities. It aims to identify the maximum-margin hyperplane that can separate the training set of points (xi, yi), where x represents the feature vector and y denotes the class. Specifically, it looks for a hyperplane that can divide the points with yi = 1 and yi = −1.

The equation of this hyperplane is as in (2)

$$w \cdot x - b = 0 \quad (2)$$

To achieve a good separation of the points, the SVM tries to maximize the margin (γ), which is subject to the constraint that for every i, γ must be less than or equal to yi(w · xi + b). This margin helps to ensure that the points are well-separated and easily classified. [2]

Moving onto logistic regression which is another model which uses mathematical relations of two factors for prediction. This algorithm predicts the probability of a certain class based on certain dependent variables. Logistic regression is a type of supervised learning technique that predicts the probability of the occurrence of an instance in a class. A binary classification issue classifier that forecasts the class probabilities. The class probabilities are output using a sigmoid function, hence the name "logistic regression" for the technique. The training algorithm employs the one-vs-rest strategy to address multiclass problems.[6][3]

TABLE 1. Accuracy Comparison of Each Algorithm

ALGORITHM	ACCURACY
Naïve Bayes	84.85%.
Support vector machine	86.64%

Logistic Regression	88.72%

Fig.2. Tweet Entry Textbox

Fig.3. Tweet Analysis

Fig.4. Tweet Analysis

IV. RESULT

Upon careful analysis using various algorithms, we found that multinomial naive Bayes showed lesser accuracy than logistic regression. Support vector machines showed lesser accuracy than logistic regression but showed better accuracy than naive Bayes. On analysis of the tweet, we get to know the sentiment of the tweet, if the said tweet is positive, neutral, or negative.

REFERENCES

[1] Natural Language Processing Elizabeth D. Liddy Syracuse University

[2] Twitter Sentiment Analysis by Vedurumudi Priyanka Sridevi Women's Engineering College, Hyderabad, India (June 13 2021)

[3] NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis Edilson A. Correa Jr., Vanessa Queiroz Marinho, Leandro Borges dos Santos ˆ Institute of Mathematics and Computer Science University of S˜ao Paulo (USP) S˜ao Carlos, S˜ao Paulo, Brazil

[4] How Will Your Tweet Be Received? Predicting the Sentiment Polarity of Tweet Replies Soroosh Tayebi Arasteh∗†‡, Mehrpad Monajem† , Vincent Christlein† , Philipp Heinrich† , Anguelos Nicolaou† , Hamidreza Naderi Boldaji† , Mahshad Lotfinia§ and Stefan Evert† †Friedrich-Alexander-Universitat Erlangen-N ¨ urnberg, Germany ¨ ‡Harvard Medical School, United States §Sharif University of Technology, Iran

[5] Twitter Sentiment Analysis Aliza Sarlan1 , Chayanit Nadam2 , Shuib Basri3 Computer Information Science Universiti Teknologi PETRONAS Perak, Malaysia

[6] Kevin P. Murphy. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press